

# Deterministic and Fast Randomized Test-and-Set in Optimal Space \*

George Giakkoupis  
INRIA Rennes  
george.giakkoupis@inria.fr

Maryam Helmi  
University of Calgary  
mhelmikh@ucalgary.ca

Lisa Higham  
University of Calgary  
higham@ucalgary.ca

Philipp Woelfel  
University of Calgary  
woelfel@ucalgary.ca

## Abstract

The test-and-set object is a fundamental synchronization primitive for shared memory systems. A test-and-set object stores a bit, initialized to 0, and supports one operation, `test&set()`, which sets the bit's value to 1 and returns its previous value. This paper studies the number of atomic registers required to implement a test-and-set object in the standard asynchronous shared memory model with  $n$  processes. The best lower bound is  $\log n - 1$  for obstruction-free [19] and deadlock-free [30] implementations. Recently a deterministic obstruction-free implementation using  $O(\sqrt{n})$  registers was presented [17]. This paper closes the gap between these known upper and lower bounds by presenting a deterministic obstruction-free implementation of a test-and-set object from  $\Theta(\log n)$  registers of size  $\Theta(\log n)$  bits.

We also provide a technique to transform any deterministic obstruction-free algorithm, in which, from any configuration, any process can finish if it runs for  $b$  steps without interference, into a randomized wait-free algorithm for the oblivious adversary, in which the expected step complexity is polynomial in  $n$  and  $b$ . This transformation allows us to combine our obstruction-free algorithm with the randomized test-and-set algorithm by Giakkoupis and Woelfel [19], to obtain a randomized wait-free test-and-set algorithm from  $\Theta(\log n)$  registers, with expected step-complexity  $\Theta(\log^* n)$  against the oblivious adversary.

---

\*The results in this paper combine and elaborate on previous work that appeared in the 2015 ACM Symposium on Theory of Computing [18] and 2013 International Symposium on Distributed Computing [17]

# 1 Introduction

A *test-and-set* (TAS) object is perhaps the simplest standard shared memory primitive that has no wait-free deterministic implementation from registers. It stores a bit, which is initially 0, and supports one operation, namely `test&set()`. A `test&set()` sets the bit's value to 1 and returns its previous value.

TAS objects have consensus number two. That is, they can be used together with registers to solve deterministic wait-free consensus only in systems with two processes. Despite that, TAS is a standard building block for shared memory algorithms that solve many classical problems, such as mutual exclusion and renaming [5–7, 12, 14, 27, 29]. Since TAS objects are among the simplest synchronization primitives, they are well suited for investigating the difficulties arising in synchronization problems. Algorithms or impossibility results for TAS provide insights into the complexity of other shared memory problems, and can contribute to their solutions.

We consider a standard shared memory system in which  $n$  processes communicate through atomic read and write operations on shared registers. A common assumption is that each register can store  $\Theta(\log n)$  bits, although in some settings registers of unbounded size are assumed. The strongest reasonable progress condition is *wait-freedom*, which guarantees that every operation finishes in a finite number of the calling process' steps, independent of other processes. Since TAS has consensus number two, deterministic wait-free implementations from registers do not exist for two or more processes. A weaker progress condition, and the one most frequently used for analyzing space complexity, is *obstruction-freedom* [24]. It guarantees that from every reachable configuration and for any process, that process will finish its operation in a finite number of its own steps, provided that no other process takes any steps (i.e., in a sufficiently long solo execution). Any shared memory object has an obstruction-free implementation from  $n$  registers [23].

The randomized step complexity of TAS has been thoroughly investigated, with significant progress being made in recent years [2, 4, 7, 20, 31]. In contrast, little was known about the space complexity of obstruction-free or randomized wait-free TAS. In 1989, Styer and Peterson [30] studied the space complexity of the related mutual exclusion problem, under the deadlock-free progress requirement. As a special case they also considered a variant called weak leader election (see Section 2). It suffices to add a single one-bit register to transform any deadlock-free weak leader election protocol into a linearizable deadlock-free TAS. Styer and Peterson proved a space lower bound of  $\lceil \log n \rceil + 1$  registers, and provided an algorithm that established that this bound is tight. Hence, in the case of deadlock-freedom, Styer and Peterson's results answer the question of the space complexity of TAS precisely up to a single register.

Deadlock-freedom is a natural progress property for mutual exclusion related problems, where waiting for other processes is inherent in the problem specification. But for other problems, it is inappropriate because it does not preclude a single slow or failing process preventing all other processes from making progress. Alternative progress properties, such as obstruction-freedom, lock-freedom, or randomized wait-freedom, are more desirable for such problems. Research on the space complexity of shared memory problems has focused on the obstruction-free progress property [16, 22, 24, 26]. However, despite significant research on TAS, prior to the result presented here, the asymptotic space complexity of obstruction-free TAS implementations remained unknown.

In 2012, Giakkoupis and Woelfel [19] used the same lower bound technique as that of Styer and Peterson to conclude that obstruction-free TAS requires  $\lceil \log n \rceil - 1$  registers. The maximum number of steps taken by any process running alone, until it finishes its method call is called the *solo step complexity* of that method [10]. In 2013 we devised a deterministic obstruction-free TAS algorithm using  $\Theta(\sqrt{n})$  registers, where the solo step complexity of `test&set()` is  $\Theta(\sqrt{n})$  [17]. We now present an asymptotically tight result.

**Theorem 1.** *There is a deterministic obstruction-free implementation of a TAS object from  $\Theta(\log n)$  registers of size  $\Theta(\log n)$  bits, where the solo step complexity of the `test&set()` method is  $\Theta(\log n)$ .*

There are performance benefits if the solo run that is required for termination is short, because processes have a better chance of completing their method call before they get interrupted. In our algorithm, processes make partial progress even if they can run uninterruptedly for a constant number of steps. As a result, a process needs to execute only a constant number of solo steps  $\Theta(\log n)$  times, to finish its `test&set()` method call.

The relation between wait-freedom and obstruction-freedom has been investigated before: Fich, Luchangco, Moir, and Shavit [16] showed that obstruction-free algorithms can be transformed into wait-free ones in the unknown-bound semi-synchronous model. The approach in this paper is different; we use randomization, but stay in the fully asynchronous model. It is easy to see that any deterministic obstruction-free algorithm can be transformed into an algorithm that is randomized wait-free against the oblivious adversary and has exponential expected step complexity. In Section 6, we provide a more efficient but also simple transformation to show the following result.

**Theorem 2.** *Suppose there is a deterministic obstruction-free algorithm whose solo step complexity is  $b$ . Then the algorithm can be transformed into a randomized one that uses the same number of registers of the same size, such that for any schedule determined by an oblivious adversary, each process finishes after at most  $O(b(n + b) \log(n/\delta))$  of its own steps with probability at least  $1 - \delta$ , for any  $\delta > 0$  (which can be a function of  $n$ ).*

We apply this transformation to our obstruction-free algorithm and combine the result with the test-and-set algorithm by Giakkoupis and Woelfel [19], to obtain a randomized wait-free TAS implementation from  $\Theta(\log n)$  registers, which has expected step complexity  $O(\log^* n)$ .

**Theorem 3.** *There is a randomized TAS implementation from  $\Theta(\log n)$  registers of size  $\Theta(\log n)$  bits, such that for any schedule determined by an oblivious adversary, the maximum number of steps executed by any process is  $O(\log^* n)$  in expectation, and  $O(\log n)$  with high probability,  $1 - n^{-\Omega(1)}$ .*

A long-lived test-and set object provides an operation `reset()` in addition to `test&set()`. The `reset()` operation can only be executed by a process if its preceding operation on the object was a successful `test&set()`; in that case the `reset()` operation unconditionally resets the value of the TAS object to 0. Recently, Aghazadeh and Woelfel [3] showed that any TAS object implemented from  $m$   $\ell$ -bit registers can be transformed into a long-lived TAS object, using  $O(m \cdot n)$  registers of size  $\max\{\ell, \log(n + m)\} + O(1)$  bits. A `reset()` operation takes only constant time in the worst-case, and the step complexity of a `test&set()` operation of the long-lived object is the same (up to a constant additive term) as the one of the (one-shot) TAS object. Applying this to the result stated in Theorem 3, yields the following:

**Corollary 4.** *A long-lived TAS object can be implemented from  $O(n \log n)$  registers, each of size  $O(\log n)$  bits, such that the expected step complexity of `test&set()` is  $O(\log^* n)$  against the oblivious adversary, and the worst-case step complexity of `reset()` is  $O(1)$ .*

The space lower bound for mutual exclusion [13] implies that any long-lived TAS implementation requires at least  $n$  registers. Aghazadeh and Woelfel [3] also gave a construction of a long-lived TAS from  $O(n)$  registers, where the expected step complexity of `test&set()` and `reset()` is  $O(\log \log n)$  against the oblivious adversary.

Our TAS algorithms rely on two components that are of independent interest; we expect they have other applications. One is an  $M$ -component snapshot object implemented from bounded registers. A  $B$ -bounded  $M$ -component snapshot object maintains a collection of  $M$  components. Each

component stores a value of size at most  $B$  bits. The object supports two operations `update( $i, x$ )` and `scan()`. Operation `scan()` returns the values of all components, and `update( $i, x$ )` writes  $x$  to the  $i$ -th component where  $x$  has size at most  $B$  bits and  $i \in \{1, \dots, M\}$ . If each component has unbounded size, then it is simply called an  $M$ -component snapshot object. The snapshot object is an important and well-studied primitive in distributed computing. There are many implementations of snapshot objects from registers in the literature [1, 8, 9, 11, 15]. The lower bound by Jayanti, Tan and Toueg for the general class of perturbable objects implies that any implementation of an  $M$ -component snapshot object from historyless and resettable consensus objects requires at least  $M - 1$  objects, and each `scan()` operation takes at least  $M - 1$  steps [26]. Fatourou, Fich and Ruppert improved the space lower bound for  $M$ -component snapshot objects to  $M$  for implementations from registers [15]. They also showed that this lower bound is tight by providing a wait-free implementation of an  $M$ -component snapshot object from  $M$  unbounded registers. But for our test-and-set implementation we need an asymptotically optimally space efficient snapshot object that uses only bounded registers. Section 5 contains our simple obstruction-free implementation of a  $B$ -bounded  $M$ -component snapshot object from  $M + 1$  bounded registers.

**Theorem 5.** *There is an obstruction-free implementation of a  $B$ -bounded  $M$ -component snapshot object from  $M + 1$  registers of size  $\Theta(B + \log n)$  bits, where the solo step complexity of `scan()` is  $O(M)$  and the solo step complexity of `update()` is  $O(1)$ .*

The key component of our TAS algorithm is a sifter object. An  $f(k)$ -sifter, where  $f$  is a function such that  $1 \leq f(k) \leq \max\{k - 1, 1\}$  for any integer  $k \geq 1$ , supports only one method, `compete()`, which returns `win` or `lose`. In any execution where  $k$  processes call `compete()`, at most  $f(k)$  of them return `win`, and at most  $k - 1$  return `lose`. Recent randomized TAS constructions [4, 20] are based on randomized sifters, where the number of winning processes is at most  $f(k)$  in *expectation*. Here, however, we use deterministic sifters, where  $f(k)$  is a worst-case bound. Section 4 contains our sifter implementation, which establishes the following theorem.

**Theorem 6.** *There is an obstruction-free implementation of a  $\lfloor \frac{2k+1}{3} \rfloor$ -sifter from a  $\lceil 4 \cdot \log n \rceil$ -bounded 6-component snapshot object.*

By combining  $O(\log n)$  sifters, and our snapshot object from Theorem 5 we obtain our TAS implementation using  $O(\log n)$  registers.

Section 2 defines our model of computation and communication. We assume these sifter and snapshot tools to implement our deterministic TAS object in Section 3, and, together with Theorem 2, our randomized TAS object in Section 7.

## 2 Model and Preliminaries

Our model of computation and communication is the standard asynchronous shared memory model where a set  $\mathcal{P}$  of  $n$  processes with distinct identifiers communicate through shared multi-reader multi-writer registers. Each register supports two atomic operations, read and write.

An algorithm is an assignment of a program to each process. Each process' program can access that process' local registers as well as the shared registers. At each *step* by a process, that process executes a single shared memory access (or, initially, its program invocation) followed by all its subsequent local operations and random choices, up to the point where that process is poised to execute its next shared memory operation. A *schedule* is a sequence of process identifiers. A schedule,  $\sigma$ , gives rise to a sequence of steps, called an *execution* as follows. The  $i$ -th step in the execution is the next step in the program of the  $i$ -th process in  $\sigma$ .

An algorithm is *deterministic* if each process' program is deterministic. A deterministic implementation of a method is *wait-free* if, from any point of an execution and for any process, the process completes its method call in a finite number of its own steps, regardless of the intervening steps taken by other processes. A deterministic implementation of a method is *obstruction-free* if, from any point of an execution and for any process  $p$ ,  $p$  completes its method call in a finite number of its own steps, provided there are no intervening steps taken by other processes. In such an execution, we say that  $p$  runs *solo* during these uninterrupted steps by  $p$ .

The algorithm is *randomized* if some process' program is randomized. An implementation of a method is *randomized wait-free* if, from any point of an execution and for any process  $p$ , the number of steps by  $p$  required for  $p$  to complete its method call is finite in expectation, regardless of the intervening steps taken by other processes [23].

A *test-and-set* (TAS) object stores one bit, which is initially 0, and supports a `test&set()` operation that sets the bit's value to 1 and returns its previous value.

An  $f(k)$ -*sifter* object, where  $f$  is a function such that  $1 \leq f(k) \leq \max\{k-1, 1\}$  for any integer  $k \geq 1$ , supports only one operation, `compete()`, which returns `win` or `lose`. In any execution where  $k$  processes call `compete()`, at most  $f(k)$  of them return `win`, and at most  $k-1$  return `lose`.

A  $B$ -bounded  $M$ -component snapshot object stores a vector  $V = (V_1, \dots, V_M)$  of  $M$  values from some domain  $D$ , where each  $d$  in  $D$  has size at most  $B$  bits. It supports two operations: `scan()` takes no parameter and returns the value of  $V$ , and `update(i, x)`,  $i \in \{1, \dots, M\}$ ,  $x \in D$ , writes  $x$  to the  $i$ -th component of  $V$  and returns nothing.

An object is *implemented* by providing a program, the *method* for `op`, for each operation, `op`, defined for that object. Since our objective is to implement a TAS object, we need to provide a `test&set` method. Our TAS algorithm is then just the `test&set` method assigned to each process. Our correctness condition is *linearizability* [25], which requires that for any execution of our algorithm and for every `test&set` method call,  $\mu$ , in that execution, there is a point between  $\mu$ 's invocation and response such that if the entire method call is replaced by the atomic `test&set` operation returning the same value as  $\mu$  at that point, the resulting execution is valid for the TAS object. Linearizability is a *composable* property: A linearizable implementation of object  $A$  assuming atomic objects  $B$ , composed with a linearizable implementation of  $B$  assuming atomic objects  $C$ , is a linearizable implementation of  $A$  using  $C$ . We exploit this by providing a linearizable implementation of a TAS object assuming an  $M$ -component snapshot object, and then a linearizable implementation of an  $M$ -component snapshot on registers. Linearizability is also a *local* property: any correct deterministic algorithm that uses a collection of atomic objects, will remain correct if these objects are replaced with their linearizable implementations.

Implementing a TAS object is related to solving *weak leader election*, where each participating process has to decide on one value, `win` or `lose`. Among all processes that finish their weak leader election protocol, at most one process is allowed to win, and not all processes may lose. Hence, if all processes finish, then exactly one process, the leader, wins. (The term leader election is ambiguous. It is also used to denote the *name consensus* problem, where the losing processes need to output the ID of the winner. We add the qualifier “weak” in order to distinguish the two variants.) Weak leader election and test-and-set are equally hard problems with respect to asymptotic space complexity. Replacing the return values 0 and 1 of a `test&set()` operation with `win` and `lose`, respectively, yields a weak leader election protocol. The difference is that TAS requires that the `test&set` method that returns 1 must be linearized before those that return 0, whereas weak leader election lacks the corresponding requirement for `win` and `lose`. Nevertheless, Golab, Hendler and Woelfel [21] gave an implementation of a TAS object using weak leader election and one additional register:

**Theorem 7.** [21] *A linearizable TAS object can be implemented using a weak leader election protocol and one additional multi-reader/multi-writer binary register, such that a `test&set()` method requires only a constant number of read and write operations in addition to the weak leader election protocol.*

For a deterministic obstruction-free implementation, the *solo step complexity* is the worst case over all processes  $p$  and all reachable configurations  $C$  of the number of steps taken in a solo execution by  $p$  starting at  $C$  until  $p$  terminates its method.

### 3 Space Efficient Deterministic Test-and-Set

Because of Theorem 7, to establish Theorem 1, it suffices to give an implementation of weak leader election that achieves the space and step complexity claimed in that theorem. We now describe this implementation, assuming we have the use of the sifter object of Theorem 6 and the snapshot object of Theorem 5.

An  $f(k)$ -sifter and a  $g(k)$ -sifter can be combined to obtain an  $f(g(k))$ -sifter, by letting the losers of the  $g(k)$ -sifter lose, and the winners call `compete()` on the  $f(k)$ -sifter. Hence, by combining enough sifter objects, we can obtain a 1-sifter, which is a weak leader election protocol.

In Section 4 we show how to implement a single  $\lfloor \frac{2k+1}{3} \rfloor$ -sifter from a 6-component snapshot object. The implementation is obstruction-free. Moreover, whenever a process starts running alone, it terminates after  $O(1)$  scan and write operations. By Theorem 5, we can implement a 6-component snapshot object from 7 registers, where the solo step complexity of each method is constant. Hence, using the obstruction-free snapshot implementation from Theorem 5, our  $\lfloor \frac{2k+1}{3} \rfloor$ -sifter implementation has constant solo step complexity and uses 7 registers.

Since multiple sifters are combined to construct our weak leader election algorithm (and hence our TAS implementation) it is more space efficient to replace the individual snapshot objects with a single snapshot object shared by all sifters. We can simulate  $\ell$  distinct 6-component snapshot objects by one  $(6\ell)$ -component snapshot object. By Theorem 5, we can implement such a snapshot object using  $6\ell + 1$  registers where the solo step complexity is  $O(\ell)$ . Hence, Theorem 5 and Theorem 6 combine to yield:

**Corollary 8.** *There is an obstruction-free implementation of  $\ell$  instances of  $\lfloor \frac{2k+1}{3} \rfloor$ -sifters using  $6\ell + 1$  registers, each of size  $\Theta(\log n)$ -bits, such that the solo step complexity of `compete()` is  $O(\ell)$ .*

We can implement a weak leader election protocol using a sequence of at most  $\ell = \lceil \log_{3/2} n \rceil + 1$  instances of a  $\lfloor \frac{2k+1}{3} \rfloor$ -sifter. As describe earlier, each process starts by invoking the `compete()` method of the first sifter; the winners of the  $i$ -th sifter proceed to the  $(i + 1)$ -th sifter, while the losers lose the weak leader election; the winner of the weak leader election is the process that wins the last sifter. We need to show that  $\ell$  repeated applications of function  $f(k) = \lfloor \frac{2k+1}{3} \rfloor$  to an initial value of  $k = n$  yield a value of 1.

**Lemma 9.** *Let  $f(n) = \lfloor \frac{2n+1}{3} \rfloor$ . Let  $f^{(0)}(n) = n$  and  $f^{(i+1)}(n) = f(f^{(i)}(n))$ . Then for any integer  $\ell \geq \log_{3/2} n$ ,  $f^{(\ell)}(n) = 1$  for any  $n \geq 1$ .*

*Proof.* First, observe that if  $n \geq 1$  then  $\lfloor \frac{2n+1}{3} \rfloor \geq 1$ , so  $f^k(n)$  never drops below 1 for any  $k$ . Now, we show by induction on  $k$ , that

$$f^{(k)}(n) \leq \left(\frac{2}{3}\right)^k n + 1 - \left(\frac{2}{3}\right)^k \text{ for } k \geq 0.$$

For the basis,  $k = 0$ , observe that  $f^{(0)}(n) = n = (\frac{2}{3})^0 \cdot n + 1 - (\frac{2}{3})^0$ .

For the inductive step:

$$\begin{aligned} f^{(k+1)}(n) &= f(f^{(k)}(n)) = \left\lfloor \frac{2(f^{(k)}(n)) + 1}{3} \right\rfloor \leq \frac{2(f^{(k)}(n)) + 1}{3} \\ &\leq \frac{2((\frac{2}{3})^k n + 1 - (\frac{2}{3})^k) + 1}{3} \text{ by the induction hypothesis} \\ &= \left(\frac{2}{3}\right)^{k+1} n + 1 - \left(\frac{2}{3}\right)^{k+1}. \end{aligned}$$

Thus, for any integer  $\ell \geq \log_{3/2} n$ ,  $f^{(\ell)}(n) \leq 1 + 1 - \frac{1}{n} < 2$ .

But  $f$  takes only integer values and  $\ell$  is an integer, implying that after  $\lfloor \log_{3/2} n \rfloor + 1$  applications of  $f$ , the value is at most 1.  $\square$

Thus, Theorem 1 follows from Theorem 7, Corollary 8 and Lemma 9. More precisely, we have:

**Theorem 10.** *There is a deterministic obstruction-free implementation of a TAS object from  $6 \lfloor \log_{3/2} n \rfloor + 7$  registers each of size at most  $4 \log n$  bits, where the solo step complexity of the `test&set()` method is  $\Theta(\log n)$ .*

Since Corollary 8 follows from Theorem 6 and Theorem 5, it remains to prove these two theorems to complete the implementation of our deterministic TAS object. This we do in the next two sections.

## 4 Sifter Implementation

This section establishes Theorem 6. Our sifter implementation is presented in Figure 1. To aid intuition we first consider a very simple obstruction-free sifter object, implemented from a 3-component snapshot object  $A$ . Each component of  $A$  can hold one process identifier. For ease of readability, we write  $A[i].\text{write}(x)$  instead of  $A.\text{update}(i, x)$ , and call `update()` operations writes. The `scan()` operation returns a triple of process identifiers, called a *signature*. At some point in an execution, process  $p$  covers component  $i$  if it writes to component  $i$  in its next step. Each process  $p$  alternates between writing and scanning. When  $p$  writes, it writes its own identifier to a component of  $A$  that did not contain  $p$  in its preceding scan. The goal of any process,  $p$ , is to achieve a *clean-sweep* meaning that its scan returns signature  $(p, p, p)$ . In this case,  $p$  terminates with `win`. If, however, while trying for a clean-sweep,  $p$ 's scan returns a signature that contains more copies of a different identifier than it has copies of  $p$ , then  $p$  terminates with `lose`. Any process that runs alone for six steps without losing, will return `win`. Furthermore, not all processes can return `lose`. To see this, let  $w$  be the last write to  $A$  and let  $p$  be the process executing  $w$ . If process  $p$  returns `lose`, then there is a process  $q$  that occupies two positions in  $p$ 's last scan, so  $q$  cannot return `lose`. Therefore, this is an implementation of an obstruction-free sifter object.

This implementation, however, is not a very efficient sifter. Suppose that while a clean-sweep is being achieved by one process, two other processes cover two distinct components of  $A$ . Then these covering processes can over-write the clean-sweep, and be made to again cover two distinct components. Now a new process can run under the cover and achieve a clean sweep. By repeating this scenario, executions are easily created where all but one process return `win`. Also, notice that to create another winner after a clean-sweep, such an obliteration of the clean-sweep by two (or three) over-writes is also necessary.

### Shared Objects:

- $A[0, 1, 2]$  is an array of the first 3 components of a 6-component snapshot object  $U$ . Each array entry stores a value from  $\mathcal{P} \cup \{\perp\}$  and is initially  $\perp$ .
- $B[0, 1, 2]$  is an array of the second 3 components of  $U$ . Each array entry stores a pair  $(\text{id}, \text{sig})$ , where  $\text{id} \in \mathcal{P} \cup \{\perp\}$ , and  $\text{sig}$  is a triple from the set  $(\mathcal{P} \cup \{\perp\})^3$ . Initially,  $\text{id} = \perp$  and  $\text{sig} = (\perp, \perp, \perp)$ .

**Notation:** For any array  $X$  and value  $v$ , let  $\text{num}(v, X) := |\{i : X[i] = v\}|$ .

**Algorithm:** compete()

```

1 pos := 0
2 while true do
3    $A[\text{pos}].\text{write}(p)$ 
4    $a := \text{scan}(A)$ 
5   if  $\text{num}(p, a) = 3$  then return win
6   if  $\exists q \in \mathcal{P} : \text{num}(p, a) < \text{num}(q, a)$  then return lose
7   if  $\text{num}(p, a) = 1$  then
8     if knockout( $a$ ) then return lose
9   Let  $\text{pos} \in \{0, 1, 2\} : a[\text{pos}] \neq p$  and  $a[(\text{pos} - 1) \bmod 3] = p$ 

```

**Function:** knockout(sig)

```

10 index := 0
11 while true do
12    $B[\text{index}].\text{write}((p, \text{sig}))$ 
13    $(\hat{a}, \hat{b}) := \text{scan}(A, B)$ 
14   if  $\hat{a} \neq \text{sig}$  then return true
15   if  $\exists q \in \mathcal{P} : q \neq p$  and  $\text{num}((q, \text{sig}), \hat{b}) \geq 2$  then return true
16   if  $\text{num}((p, \text{sig}), \hat{b}) = 3$  then return false
17   Let  $\text{index} \in \{0, 1, 2\} : \hat{b}[\text{index}] \neq (p, \text{sig})$ 

```

Figure 1: Implementation of a sifter for process  $p \in \mathcal{P}$



To reduce the number of processes that can return **win** to at most a constant fraction of those that compete, the core idea is to prevent processes that participate in over-writing a clean-sweep, from covering again, without some process losing. This is achieved, in our algorithm, by expanding the 3-component snapshot object  $A$  with 3 additional components. The first 3 components are referred to as  $A$ , and the second 3 components as  $B$ . We implement  $A$  and  $B$  together from a 6-component snapshot object  $U$ . To make notation more intuitive we use the following convention: for each  $i \in \{0, 1, 2\}$ ,  $A[i].\text{write}(x)$  denotes  $U.\text{update}(i, x)$  and  $B[i].\text{write}(x)$  denotes  $U.\text{update}(i + 3, x)$ . Furthermore,  $\text{scan}(A, B)$  returns simply what  $U.\text{scan}()$  returns, and  $\text{scan}(A)$  returns the first three components returned by  $U.\text{scan}()$ .

Each component of  $B$  can hold a pair consisting of a process identifier and a signature. A write by  $p$  can be either a write of  $p$  to a component of  $A$ , or a write of  $(p, s_p)$  to a component of  $B$ , where  $s_p$  is a signature. Each process begins by competing on  $A$  and still strictly alternates between writing and scanning.

If process  $p$ , competing on  $A$ , gets a scan with signature  $s$  of  $A$ , where the identifiers in  $s$  are all distinct and one of them is  $p$ , then  $p$  leaves  $A$  to compete on  $B$  while remembering  $s$ . (Notice that if  $p$  does not get such a scan and it does not immediately return **lose**, then  $p$  is in at least two positions in  $s$ . Therefore, its last write could not have been part of an over-write of a clean-sweep by some other process.) By writing the pair  $(p, s)$  to components of  $B$ ,  $p$  tries to achieve a clean-sweep of  $B$  (meaning a scan by  $p$  shows that each of the 3 components of  $B$  contains  $(p, s)$ ). If  $p$  achieves such a clean-sweep, then it returns to competing on  $A$ , as described above. There are two ways that process  $p$  can lose while playing on  $B$ . First,  $p$  loses if, while trying to achieve a clean-sweep of  $B$ , one of  $p$ 's scans shows a signature of  $A$  different from  $s$ . Second,  $p$  loses if its scan shows that for some other process  $q$ ,  $(q, s)$  occupies at least 2 positions of  $B$ . That is,  $p$  only returns to continue competing on  $A$  if it achieves a clean-sweep of  $B$  while each of its scans satisfies 1) the signature of  $A$  is  $s$ , and 2) no other process with signature  $s$  occupies more than one component of  $B$ .

#### 4.1 Intuition for Correctness

Our proof will establish that not all processes can return **lose**, and at most  $\lfloor (2k + 1)/3 \rfloor$  processes can win, if  $k$  processes participate. While the proof has to attend to several subtleties and substantial detail, there are several insights that aid our intuition. We say a process is playing on  $A$ , if its next shared memory step is on  $A$ . Consider the three ways that a process can return **lose**. Let us say  $p$  *loses on  $A$*  if process  $p$  loses while playing on  $A$  because the signature of  $A$  in its last scan contained more occurrences of some other process than occurrences of  $p$ . We say  $p$  *signature-loses on  $B$*  if process  $p$  with signature  $s$ , loses while trying to achieve a clean-sweep of  $B$ , because one of  $p$ 's scans shows a signature of  $A$  different from  $s$ . We say  $p$  *process-loses on  $B$*  if process  $p$  loses because its scan shows that for some other process  $q$ ,  $(q, s)$  occupies at least 2 positions of  $B$ .

Lemma 17 below states that not all processes can lose. For the intuition suppose that all processes lose. Consider the last write, say  $w$ , to  $A$ , and let  $p$  be the process that executes  $w$ . Process  $p$  cannot lose on  $A$  because if it did, then in  $p$ 's last scan there is some process,  $q \neq p$ , that occupies 2 positions on  $A$ , and that process cannot return **lose** unless some process writes to  $A$  after  $w$ . Similarly,  $p$  cannot signature-lose on  $B$  because, again, that would imply a write to  $A$  after  $w$ . So suppose  $p$  process-loses on  $B$ . Then we show that there is some other process, say  $q$ , that has the same signature as  $p$  and is competing with  $p$ , and  $q$  cannot process-lose on  $B$ . Process  $q$  also cannot signature-lose on  $B$  or lose on  $A$  without a write to  $A$  happening after  $w$ .

Lemma 23 below states that if  $k$  processes call **compete()**, then at most  $\lfloor (2k + 1)/3 \rfloor$  of them win. Consider the intervals in an execution between the final scans of processes that return **win**

(achieve a clean-sweep of  $A$ ). If  $\ell$  processes return **win**, there are  $\ell - 1$  such disjoint intervals. We associate each such interval  $I$ , with a losing process as follows.

1. If  $I$  contains the last write by a process  $p$  that loses on  $A$ , then associate  $I$  with  $p$ .
2. If  $I$  contains the last write by a process  $q$  that signature-loses on  $B$ , associate  $I$  with  $q$ .
3. If  $I$  is not associated with a losing process via either (1) or (2), we will associate  $I$  with a losing process as follows.

We will prove that there is a sub-interval  $I'$  of  $I$  and there are either two or three processes that, during  $I'$ , move from  $A$  to  $B$  and finish competing on  $B$  using some signature, say  $s$ , while the signature of  $A$  remains  $s$  throughout  $I'$ . Now we focus on the execution during  $I'$ . Since  $B$  has three components, after any clean-sweep on  $B$ , a subsequent clean-sweep on  $B$  requires two processes to over-write the previous clean-sweep. These over-writers must have signature  $s$ , because, otherwise, an over-writer has a signature different from that of  $A$  and would signature-lose on  $B$ , implying that  $I$  has an associated losing process via (2). If there are two processes with signature  $s$  then  $I'$  can have at most one clean-sweep, and if there are three processes then  $I'$  can have at most two clean-sweeps. Therefore, at least one of the two or three processes competing on  $B$  with signature  $s$  cannot return **win**, and  $I$  is associated with one such process. Notice, however, that this process could withhold its last write in order to be assigned to a later interval via (2).

Therefore, using these three rules of association, we assign at least one losing process to every interval, and no process is assigned to more than two of these intervals. Thus there are at least  $(\ell - 1)/2$  processes that cannot return **win**.

## 4.2 Notation and Terminology

Throughout the remainder of the section we consider a fixed execution  $E$ . A *losing scan* is a scan by a process such that this process will return **lose** in its next step, without doing any further shared memory operation. A *winning scan* is a scan by a process such that this process will return **win** in its next step, without doing any further shared memory operation. For each winning scan there exists a last write by the process that performs this scan. We call this write a *winning write*. Let  $s_1, s_2, \dots, s_\kappa$  be the sequence of winning scans in  $E$  and let  $q_1, q_2, \dots, q_\kappa$  denote the corresponding sequence of processes that performed these scans. Observe that for all  $i$ ,  $1 \leq i \leq \kappa$ ,  $s_i$  is preceded by a winning write  $w_i$  performed by  $q_i$ . Furthermore,  $s_i$  must happen before  $w_{i+1}$  because at  $s_i$  all components in  $A$  contain  $q_i$ 's id however, at  $w_{i+1}$ ,  $q_{i+1}$  has written its own id everywhere in  $A$ . Hence winning scans and winning writes strictly interleave. That is, the order of winning scans and writes in  $E$  is  $w_1, s_1, w_2, s_2, \dots, w_\kappa, s_\kappa$ .

Suppose  $E = \text{op}_1, \text{op}_2 \dots$ , we denote the contiguous subsequence of  $E$  starting at  $\text{op}_i$  and ending at the operation immediately before  $\text{op}_j$  by  $E[\text{op}_i : \text{op}_j)$ . A *sifting interval* is a subsequence of an execution that starts at some winning scan and ends at the operation immediately before the next winning write. Observe that all sifting intervals are disjoint. Also because there has been a preceding winning scan, no component of  $A$  contains  $\perp$  in any sifting interval. Note that since  $E$  contains  $\kappa$  winning scans it has  $\kappa - 1$  disjoint sifting intervals.

A *signature* is an ordered triple of identifiers. A signature  $(p_0, p_1, p_2)$  is *full* if for any  $i, j \in \{0, 1, 2\}$ ,  $i \neq j$  implies  $p_i \neq p_j$ .

The following lemmas concern properties of executions. Terms such as before, after, next, previous, precedes, and follows are all with respect to the order of operations in execution  $E$ .

A local variable  $x$  in the algorithm is denoted by  $x_p$  when it is used in the method call invoked by process  $p$ .

### 4.3 Proof of Correctness

Lemmas 11 through 16 provide us with some properties of the algorithm that are used in Lemma 17, to prove that there is no execution in which all processes return **lose**.

**Lemma 11.** *Suppose that process  $p$  executes a scan, say  $s$ , at Line 13, and in this scan( $A, B$ ),  $A = \sigma$ . If  $s$  is not a losing scan then, at the most recent scan( $A$ ) executed in Line 4, by  $p$ , preceding  $s$ ,  $A = \sigma$ .*

*Proof.* Let  $\hat{s}$  be the most recent scan( $A$ ) executed in Line 4, by  $p$ , preceding  $s$ . By way of contradiction suppose that at  $\hat{s}$ ,  $A = \sigma'$ , where  $\sigma' \neq \sigma$ . Then, at  $s$ , by Line 8,  $\text{sig}_p = \sigma'$  and by Line 13,  $\hat{a}_p = \sigma$ . Hence at  $s$ ,  $\hat{a}_p \neq \text{sig}_p$ . Therefore, by Lines 14 and 8,  $s$  is a losing scan which is a contradiction.  $\square$

**Lemma 12.** *Let  $s$  be any scan by process  $p$  and  $w$  be  $p$ 's next write. If, at  $w$ ,  $p$  writes to  $A[j]$ , then at  $s$ ,  $A[j] \neq p$  and  $A[(j-1) \bmod 3] = p$ .*

*Proof.* Let  $\hat{s}$  be the last scan( $A$ ) executed in Line 4 by  $p$  preceding  $w$ . Since  $w$  is to  $A[j]$ , by Line 9, at  $\hat{s}$ ,  $A[j] \neq p$  and  $A[(j-1) \bmod 3] = p$ . Since  $p$  performs a write after  $s$ ,  $s$  is not a losing scan. By Lemma 11, the signature of  $A$  at  $s$  and  $\hat{s}$  is equal. Therefore at  $s$ ,  $A[j] \neq p$  and  $A[(j-1) \bmod 3] = p$ .  $\square$

**Lemma 13.** *Suppose at scan  $s$ ,  $A = (p_0, p_1, p_2)$  is a full signature. For any  $i \in \{0, 1, 2\}$ , if  $p_i$  writes to  $A$  after  $s$ , then its first write into  $A$  after  $s$  is not to  $A[i]$ .*

*Proof.* Let  $w_i$  be the first write by  $p_i$  to  $A$  after  $s$ . Let  $s_i$  be the scan by  $p_i$  preceding  $w_i$ . If  $s_i$  happens before  $s$ , then there is no write to  $A$  by  $p_i$  in the execution  $E[s_i : s)$ . At  $s$ ,  $A[i] = p_i$ , hence at  $s_i$ ,  $A[i] = p_i$ . Suppose  $s_i$  happens after  $s$ . At  $s$ ,  $A[i]$  is the only location that contains  $p_i$ , and there is no write to  $A$  by  $p_i$  in the execution  $E[s : s_i)$  and  $s_i$  is not a losing scan. Therefore, at  $s_i$ ,  $A[i] = p_i$ . In either case, by Lemma 12,  $w_i$  is a write to  $A[\text{pos}]$  where  $\text{pos} \neq i$ .  $\square$

**Lemma 14.** *Suppose at scan  $s$ ,  $A = (p_0, p_1, p_2)$  is a full signature. Let  $w$  be the first write to  $A$  after  $s$ . Then  $w$  changes the signature of  $A$ .*

*Proof.* Let  $q$  be the process executing  $w$ . If  $q \notin \{p_0, p_1, p_2\}$ , then since  $q$  writes its own id, it changes the signature of  $A$ . If  $q = p_i \in \{p_0, p_1, p_2\}$ , then by Lemma 13,  $q$  writes to a location different from  $A[i]$ . Hence  $w$  changes the signature of  $A$ .  $\square$

**Lemma 15.** *Suppose at scan  $s_1$ ,  $A = (p_0, p_1, p_2)$  is a full signature. Let  $w$  be the first write to  $A$  after  $s_1$ . Let  $s_2$  be any scan after  $w$  such that at  $s_2$ ,  $A = (p_0, p_1, p_2)$ . Then, for some  $\ell \in \{0, 1, 2\}$ ,  $p_\ell$  calls **knockout**( $\sigma$ ), where  $\sigma \neq (p_0, p_1, p_2)$  and returns **false** in the execution  $E[w : s_2)$ .*

*Proof.* Suppose that  $w$  is a write to component  $A[i]$ . Since  $A[i] = p_i$  at  $s_2$ , the last write to  $A[i]$  in  $E[w : s_2)$ , is by  $p_i$ .

By Lemma 13, the first write by  $p_i$  to  $A$  in  $E[w : s_2)$ , say  $w_1^i$ , is to  $A[j]$  where  $j \neq i$ . Thus  $p_i$  must perform at least two writes to  $A$  in the interval  $E[w : s_2)$ . Let  $s_2^i$  be the scan by  $p_i$  following  $w_1^i$  in  $E[w : s_2)$ , and  $\sigma$  be the signature of  $A$  at  $s_2^i$ .

Since  $w_1^i$  is to  $A[j]$  and  $w$  is to  $A[i]$ ,  $w \neq w_1^i$ , implying  $w$  is not executed by  $p_i$ . Immediately after  $w$ , no location in  $A$  contains  $p_i$ . Because  $p_i$  writes only once in  $E[w : s_2^i)$ , at  $s_2^i$ ,  $p_i$  can appear only in  $A[j]$ . Since  $s_2^i$  is not a losing scan,  $p_i$  must still be in  $A[j]$  at  $s_2^i$ , and  $\sigma$  be full. This implies  $p_i$  calls **knockout**( $\sigma$ ) after  $s_2^i$ . Furthermore,  $\sigma \neq (p_0, p_1, p_2)$ . Finally, because  $p_i$  writes to  $A$  after  $s_2^i$ ,  $p_i$  must return **false** from this **knockout** call.  $\square$

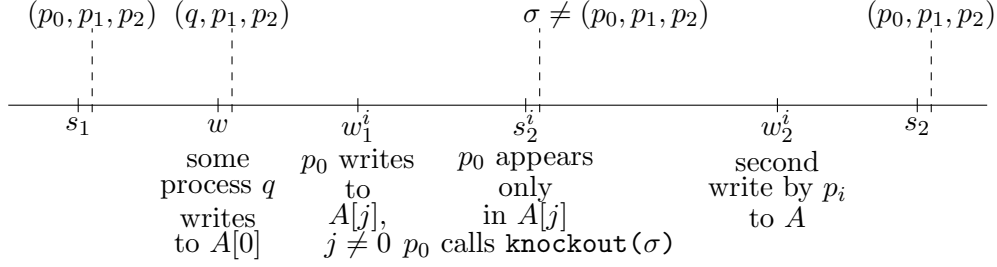


Figure 2: Illustration of the order of the operations, where  $i = 0$

Recall that in any execution, if a process  $p$  performs a  $\text{scan}(A)$  in which  $A = \sigma$  is a full signature containing  $p$ , it invokes  $\text{knockout}(\sigma)$ . During this  $\text{knockout}(\sigma)$  call,  $p$  tries to write  $(p, \sigma)$  to all components of  $B$ . Let  $w$  be any write of this  $\text{knockout}(\sigma)$  call. In the following lemma, we prove that if at some scan after  $w$ , say  $s$ ,  $A = \sigma$  and  $B[i] = (p, \sigma)$ , then the signature of  $A$  is  $\sigma$  in the entire execution between  $w$  and  $s$ . In other words, during  $E[w : s]$ , the signature of  $A$  cannot change from  $\sigma$  to  $\sigma' \neq \sigma$  and change back to  $\sigma$  again while  $p$  is performing one single  $\text{knockout}(\sigma)$ .

**Lemma 16.** *Suppose at scan  $s$ ,  $A = \sigma$  is a full signature, and there is an  $i \in \{0, 1, 2\}$  and a process  $p$  such that  $B[i] = (p, \sigma)$ . Let  $w_i$  be the last write to  $B[i]$  that precedes  $s$ . Then, there is no write to  $A$  in  $E[w_i : s]$ .*

*Proof.* By way of contradiction, let  $w$  be the first write to  $A$  in  $E[w_i : s]$ . Let  $s_i$  be the last scan by  $p$  preceding  $w_i$ . Since  $w_i$  has value  $(p, \sigma)$ , at  $w_i$ ,  $\text{sig}_p = \sigma$ . Therefore, at the last scan executed in Line 4 preceding  $w_i$ ,  $A = \sigma$ .

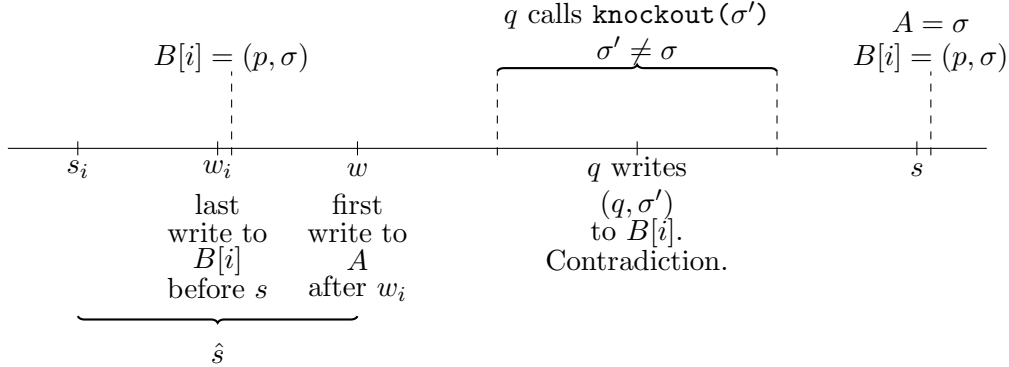


Figure 3: Illustration of the order of the operations if  $w_i$  precedes  $\hat{w}$

Hence, by Lemma 11, at  $s_i$ ,  $A$  must have signature  $\sigma$ . Let  $\hat{s}$  be the last scan before  $w$  in which the signature of  $A$  is  $\sigma$ . Since  $s_i$  precedes  $w$ ,  $\hat{s}$  exists. Then  $w$  is the first write to  $A$  following  $\hat{s}$ . By Lemma 15, there is a process  $q$  that executes a complete  $\text{knockout}(\sigma')$  in  $E[w : s]$ , where  $\sigma \neq \sigma'$ , and returns **false**. Hence, in  $E[w_i : s]$ ,  $q$  over-writes every component in  $B$  with  $(q, \sigma')$ . This contradicts that  $w_i$  is the last write to  $B[i]$  preceding  $s$ .  $\square$

**Lemma 17.** *There is no execution in which all processes return **lose**.*

*Proof.* By way of contradiction, assume that there is an execution in which all processes return **lose**. Let  $u$  be the process that performs the last write to  $A$ , let  $w_u^A$  be that write, and let  $\sigma$  be the signature of  $A$  after  $w_u^A$ . Let  $s_u$  be the last scan by  $u$ . Then  $u$  returns **lose** in Line 6 or 8.

First consider the case in which  $u$  returns **lose** in Line 6. At  $s_u$ ,  $\text{num}(u, a_u)$  is not equal to 0 because the last write to  $A$  is performed by  $u$  and  $s_u$  happens after  $w_u^A$ . Therefore, by the if-condition of Line 6,  $\text{num}(u, a_u) = 1$  and there is a process  $y$  such that in  $s_u$ ,  $\text{num}(y, a_u) = 2$ . Let  $w_y^A$  be the last write by  $y$  to  $A$ . Since  $u$  performs the last write to  $A$ ,  $w_y^A$  precedes  $w_u^A$ . Because no process writes  $y$  to  $A$  after  $w_y^A$  and no process writes to  $A$  after  $w_u^A$  and, later, at  $s_u$ ,  $\text{num}(y, a_u) = 2$ , it follows that  $\text{num}(y, A) \geq 2$  for the entire execution after  $w_y^A$ . Therefore any scan by  $y$  after  $w_y^A$  must satisfy  $\text{num}(y, a_y) \geq 2$ . This implies  $y$  cannot return **lose**, contradicting the assumption.

Next consider the case in which  $u$  returns **lose** in Line 8. This implies  $u$  calls **knockout**( $\sigma$ ) after  $w_u^A$  from which it returns **true** in Line 14 or in Line 15. But  $u$  cannot return **true** in Line 14 because the value of array  $A$  remains  $\sigma$  after  $w_u^A$ . Therefore  $u$  returns **true** in Line 15.

Let  $S = \{(q, i, w_q) \mid B[i] = (q, \sigma) \text{ at some scan after } w_u^A \text{ and } w_q \text{ is the last write by } q \text{ to } B[i] \text{ before this scan}\}$  and let  $Q = \{q \mid (q, i, w_q) \in S\}$ . Because  $u$  returns **true** in Line 15,  $S$  is not empty. By Lemma 16, for each  $(q, i, w_q) \in S$ ,  $w_u^A$  precedes  $w_q$ . This implies that for each  $q \in Q$ ,  $q$  performs a write (i.e.  $w_q$ ) to  $B$  after  $w_u^A$  and, by assumption, some time later, does a losing scan.

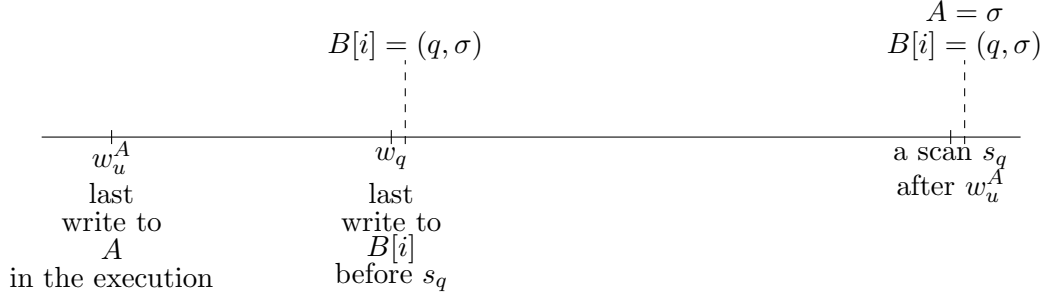


Figure 4: Illustration of the order of the operations if  $w_i$  precedes  $\hat{w}$

For each  $q \in Q$ ,  $q$  cannot return **lose** at Line 6 because this would imply  $q$  writes to  $A$  after  $w_u^A$ . Therefore  $q$  returns **lose** at Line 8 implying that  $q$  returns **true** at Line 14 or Line 15. It does not return **true** in Line 14 because  $\text{sig}_q = \sigma$  and the value of  $A$  remains  $\sigma$  after  $w_u^A$ . Therefore for each  $q \in Q$ ,  $q$  returns **true** in Line 15, following a losing scan that is after  $w_u^A$ . Let  $z$  be the last process in  $Q$  to do its losing scan,  $s_z$ . At  $s_z$  two components in  $B$  contain  $(z', \sigma)$ , where  $z' \neq z$ . Hence,  $z' \in Q$ . Thus, between the last write by  $z'$  (after  $w_u^A$ ) and the last scan by  $z'$ , these two components in  $B$  contain  $(z', \sigma)$ . So at  $z'$ 's last scan at least two components in  $B$  contain  $(z', \sigma)$ . Therefore  $z'$  cannot return **true** in Line 15, contradicting the assumption that the last scan of  $z'$  is a losing scan.  $\square$

Lemma 18 through Lemma 24 provide us with additional properties of the algorithm that are combined to prove, in Lemma 25, that when two or more processes invoke **compete**(), at most a constant fraction of them can return **win**.

**Lemma 18.** *Suppose an execution between a write and the next scan by the same process, say  $p$ , contains a winning write. Then the scan by  $p$  is a losing scan.*

*Proof.* At the winning write all components in  $A$  contain the id of the process that performs this winning write. In the sub-execution from the winning write to the scan by  $p$  there is no write by  $p$ . Since only  $p$  writes its id, at  $p$ 's scan,  $\text{num}(p, a_p) = 0$ . Hence  $p$  returns **lose** after this scan.  $\square$

**Observation 19.** *Every sifting interval contains at least two writes to  $A$ .*

*Proof.* Consider the sifting interval  $I = E[s^* : w^*]$ . Let  $p^*$  be the process that performs  $w^*$ . Since at the winning write  $w^*$ , all components in  $A$  contain  $p^*$ ,  $p^*$  must have performed at least two writes to  $A$  before  $w^*$ , and these two writes must be after the previous winning scan, which is  $s^*$ .  $\square$

A sifting interval that does not contain a write to  $A$  by a process whose next scan is a losing scan is called a *slow sifting interval*.

**Lemma 20.** *For any slow sifting interval  $I$ , there exists a signature  $\sigma = (q_0, q_1, q_2)$  and a set  $Z \subseteq \{0, 1, 2\}$  satisfying:  $|Z| = 2$  and for each  $z \in Z$  during  $I$ ,  $q_z$  performs a write and then a scan in `compete()` and then invokes `knockout( $\sigma$ )` and becomes poised to write  $(q_z, \sigma)$  to  $B$ . Furthermore, there is no write to  $A$  between these two scans.*

*Proof.* Let  $I$  be  $E[s^* : w^*]$  and Let  $p^*$  be the process that performs  $s^*$ . Suppose that  $w_1^A, w_2^A, \dots, w_\ell^A$  is the sequence of all writes to  $A$  during  $I$ . By Observation 19,  $\ell \geq 2$ . For each  $i$ ,  $1 \leq i \leq \ell$ , let  $s_i$  denote the next scan by the process that executes  $w_i^A$ . Each  $s_i$  is at Line 4 following  $w_i^A$ , at Line 3 of `compete()`. Let  $S$  denote the set of all these scans. Let  $\hat{I}$  denote the execution  $E[w_2^A : w^*]$ .

By Lemma 18, if  $s_i$  happens after  $w^*$  then  $s_i$  is a losing scan and hence  $E[s^* : w^*]$  is not a slow sifting interval. Therefore for all  $i$ ,  $1 \leq i \leq \ell$ ,  $s_i$  occurs in  $E[w_i^A : w^*]$ . Let  $q$  be the process that performs  $s_1$ . At  $s^*$ ,  $\text{num}(p^*, A) = 3$ . Because only one write happens to  $A$  during  $E[s^* : w_2^A]$ ,  $q$  would return `lose` at Line 6 if  $s_1$  precedes  $w_2^A$  implying  $E[s^* : w^*]$  is not a slow sifting interval. Hence for all  $i$ ,  $1 \leq i \leq \ell$ ,  $s_i$  must happen in  $\hat{I}$ . Execution  $\hat{I}$  consists of the  $\ell - 1$  disjoint sub-executions  $E[w_2^A : w_3^A], E[w_3^A : w_4^A], \dots, E[w_\ell^A : w^* = w_{\ell+1}^A]$ . Since  $\ell$  scans happen in these  $\ell - 1$  executions, by the pigeonhole principal, there is a  $j$ ,  $2 \leq j \leq \ell$  such that (at least) two scans in  $S$ , say  $s'$  and  $s''$  occur in  $E[w_j^A : w_{j+1}^A]$ . Because no process performs two scans in `compete()` without writing to  $A$  in between,  $s'$  and  $s''$  are performed by two distinct processes say  $q_z$  and  $q_{z'}$ . Because  $E[s^* : w^*]$  is a slow sifting interval, neither  $q_z$  nor  $q_{z'}$  return `lose` at Line 6. Since no write happens to  $A$  during  $E[w_j^A, w_{j+1}^A]$ , the scans by  $q_z$  and  $q_{z'}$  in `compete()` return the same signature for  $A$ , say,  $\sigma$  where  $\sigma$  contains  $q_z$  and  $q_{z'}$ . Therefore  $q_z$  and  $q_{z'}$  both invoke `knockout( $\sigma$ )`.  $\square$

**Lemma 21.** *Suppose at scan  $s_1$ ,  $A = (p_0, p_1, p_2)$  is a full signature. Let  $w$  be the first write to  $A$  after  $s_1$ . Let  $s_2$  be any scan after  $w$  such that at  $s_2$ ,  $A = (p_0, p_1, p_2)$ . Then, for all  $\ell \in \{0, 1, 2\}$ ,  $p_\ell$  performs at least two writes to  $A$  in the execution  $E[w : s_2]$ .*

*Proof.* In order to prove the lemma we show that for each  $\ell \in \{0, 1, 2\}$ , in the execution  $E[w : s_2]$ :

- a) if  $A[\ell]$  is written, then in this execution,  $p_\ell$  writes to  $A$  at least twice;
- b)  $A[\ell]$  is written.

First we prove (a). Let  $w_\ell$  be the last write to  $A[\ell]$  during  $E[w : s_2]$ . Since  $A[\ell] = p_\ell$  at  $s_2$ ,  $w_\ell$  is executed by  $p_\ell$ . By Lemma 14,  $p_\ell$ 's first write during  $E[w : s_2]$  is not to  $A[\ell]$ . Hence,  $p_\ell$  executes at least two writes during  $E[w : s_2]$ , proving (a).

We now prove (b). Suppose that  $w$  is a write to component  $A[i]$ . By (a),  $p_i$  writes to  $A$  at least twice during  $E[w : s_2]$ . Let  $w_i^A$  be  $p_i$ 's first write to  $A$  during  $E[w : s_2]$ . By Lemma 13,  $w_i^A$  is to  $A[j]$  where  $j \neq i$ .

By way of contradiction assume there is a  $k \in \{0, 1, 2\}$  such that,  $A[k]$  is not written in  $E[w : s_2]$ . In particular, since  $A[i]$  and  $A[j]$  are written (by writes  $w$  and  $w_i^A$ , respectively) in  $E[w : s_2]$ , we have:

$$k \notin \{i, j\}. \quad (1)$$

Let  $w_{i'}$  and  $w_{j'}$  be the last writes to  $A[i]$  and  $A[j]$ , respectively, during  $E[w : s_2]$ . Since  $A[i] = p_i$  and  $A[j] = p_j$  at  $s_2$ ,  $w_{i'}$  is executed by  $p_i$  and  $w_{j'}$  by  $p_j$ . Let  $s_{i'}$  and  $s_{j'}$  be the scans by  $p_i$ , respectively  $p_j$ , preceding  $w_{i'}$ , respectively  $w_{j'}$ . By (a), both processes execute at least two writes during  $E[w : s_2]$ , and thus  $s_{i'}$  and  $s_{j'}$  are both also in  $E[w : s_2]$ . From Lemma 12 we conclude that  $A[(i-1) \bmod 3] = p_i$  at  $s_{i'}$  and  $A[(j-1) \bmod 3] = p_j$  at  $s_{j'}$ . Since no process writes to  $A[k]$  in  $E[w : s_2]$ ,  $A[k] = p_k$  throughout  $E[w : s_2]$ . Hence, we have:

$$k \neq (i-1) \bmod 3 \text{ and } k \neq (j-1) \bmod 3. \quad (2)$$

Combining conditions (4.1) and (4.2) contradicts that  $k$  is in  $\{0, 1, 2\}$ .  $\square$

**Lemma 22.** *Let  $s$  be a scan( $A$ ) from Line 4 by  $p$  immediately before  $p$  invokes `knockout( $\sigma$ )` and  $s'$  be any scan( $A, B$ ) by  $p$  within this invocation. Let  $w$  be the first write to  $A$  after  $s$ . If  $w$  precedes  $s'$ , then  $s'$  is a losing scan.*

*Proof.* Since  $p$  invokes `knockout( $\sigma$ )`,  $\sigma$  is a full signature,  $\text{num}(p, \sigma) = 1$  and  $A = \sigma$  at  $s$ . By way of contradiction suppose  $s'$  is not a losing scan. Hence, by Lemma 11, at  $s'$ , the signature of  $A$  is  $\sigma$ . Therefore, by Lemma 21,  $p$  writes to  $A$  in the execution  $E[w : s']$ . This is a contradiction because  $p$  is performing `knockout( $\sigma$ )` in this entire execution and there are no writes to  $A$  during the `knockout` method call.  $\square$

**Lemma 23.** *For every slow sifting interval  $I$ , there is a process  $p$  that performs a write during  $I$  and either the first or the second scan by  $p$  following this write is a losing scan.*

*Proof.* Let  $I = E[s^* : w^*]$  be a slow sifting interval. By Lemma 20, there exists a full signature  $\sigma = (q_0, q_1, q_2)$ , a set  $Q \subseteq \{q_0, q_1, q_2\}$ , satisfying  $|Q| = 2$  and for each  $q \in Q$  during  $I$ :

- 1)  $q$  performs a write to  $A$  and a scan in `compete()` and calls `knockout( $\sigma$ )` and becomes poised, at Line 12, to write  $(q, \sigma)$  to  $B[0]$ ; and
- 2) there is no write to  $A$  between these scans.

Let  $Q' \subseteq \{q_0, q_1, q_2\}$  be the set of all processes satisfying (1) and (2). Therefore  $2 \leq |Q'| \leq 3$ . Let  $\hat{s}$  be the earliest of these scans (by processes in  $Q'$  immediately before calling `knockout( $\sigma$ )`). At  $\hat{s}$ , the signature in  $A$  is full and at  $w^*$  the same id is in all locations of  $A$ . Therefore,  $w^*$  is the second or later write after  $\hat{s}$ . Hence, there is at least one write to  $A$  in  $E[\hat{s} : w^*]$ . Let  $w$  be the first write to  $A$  in  $E[\hat{s} : w^*]$ .

Suppose there is  $q \in Q'$ , such that  $q$  performs a scan, say  $s$ , in Line 13 of its current call to `knockout` after  $w$ . Then by Lemma 22,  $s$  is a losing scan. Since  $q$  writes at least once in  $E[s^* : w^*]$  and at most once after  $w$ , it follows that  $q$  performs its last or second last write during  $E[s^* : w^*]$ , and so  $s$  is either  $q$ 's first or second scan following this write, and the lemma holds.

Otherwise, all processes in  $Q'$  execute at least one write and perform their last scan of their current call to `knockout` before  $w$ . We partition this case into three subcases.

Case 1: There is  $q \in Q'$  such that  $q$  calls `knockout( $\sigma$ )` and returns `true` (Line 14 or 15). Then  $q$ 's last scan before returning `true` is a losing scan, and the lemma follows.

Case 2: For each process  $q \in Q'$ ,  $q$ 's current `knockout` call returns `false` and there is a process  $p \notin Q'$  that performs a write  $w_p$  to  $B$  with value  $(p, \sigma')$  in the execution  $E[\hat{s} : w]$  where  $\sigma' \neq \sigma$ . When  $p$  did its scan in `compete()` just before invoking `knockout( $\sigma'$ )`, the signature of  $A$  was  $\sigma'$ . At  $w_p$ , the signature of  $A$  is  $\sigma \neq \sigma'$ , so there is a write to  $A$  between this scan by  $p$  and  $w_p$ . Hence, by Lemma 22,  $p$ 's next scan after  $w_p$  is a losing scan, and again the lemma follows.

Case 3: For each process  $q \in Q'$ ,  $q$ 's current **knockout** call returns **false** and there is no write to  $B$  in  $E[\hat{s} : w)$  that contains a signature different from  $\sigma$ . We show that this case is impossible. Let  $S$  be the set of last scans of **knockout** calls by processes in  $Q'$ . Let  $s''$  be the last scan and  $s'$  be the second last scan in set  $S$ . Let  $q'$  and  $q''$  be the processes performing  $s'$  and  $s''$  respectively. Since  $q'$  returns **false**, all three components in  $B$  contain  $(q', \sigma)$  at  $s'$ . After  $s'$ , there can be at most one write to  $B$  by  $q''$ . Because  $q''$ 's next scan after such a write would be a losing scan, contradicting that  $q''$  returns **false**.  $\square$

**Lemma 24.** *For every sifting interval, there is a process  $p$  and a write  $w$  by  $p$  satisfying: either the first operation by  $p$  or the third operation by  $p$  that follows  $w$  is a losing scan.*

*Proof.* For any sifting interval that is not slow, the lemma holds by definition. For any slow sifting interval, the lemma follows from Lemma 23, because each process alternates between writes and scans.  $\square$

**Lemma 25.** *If  $k$  processes invoke the **compete()** method, then at most  $\lfloor \frac{2k+1}{3} \rfloor$  processes return **win**.*

*Proof.* If  $k'$  processes return **win**, then by definition, there are  $k' - 1$  sifting intervals. By Lemma 23, for each sifting interval there is a process that performs its last or second last write and it cannot return **win**. Hence there are at least  $\lfloor \frac{k'-1}{2} \rfloor$  processes which have invoked **compete()** and cannot return **win**. Since  $\lfloor \frac{k'-1}{2} \rfloor + k' \leq k$ ,  $k'$  is at most  $\lfloor \frac{2k+1}{3} \rfloor$ .  $\square$

**Lemma 26.** *The sifter implementation in Figure 1 is obstruction-free where each process terminates in  $O(1)$  solo steps.*

*Proof.* Suppose a process,  $p$ , begins a solo run while it is executing **knockout**. If it returns **true** in either Line 14 or Line 15, then it terminates due to Line 8. Otherwise in each iteration of the while loop, it writes a new location in  $B$ . Therefore after three iterations, all locations in  $B$  contain  $(p, \text{sig}_p)$ , and  $p$  returns **false** in Line 16. When  $p$  executes **knockout** during its solo run, the value of  $A$  is equal to  $\text{sig}_p$  because otherwise  $p$  returns **true** from its **knockout** call. In  $\text{sig}_p$ , exactly one location in  $A$  contains  $p$  and no other process writes to  $A$  after it returns from its **knockout** call. Hence  $p$  writes two more times to  $A$  and, by Line 5 returns **win**.

Suppose  $p$  starts its solo run in a **compete()** call. After at most one write it performs a scan. Then it either returns **win** due to Line 5 or returns **lose** in Line 6, or it invokes a **knockout** call. If it calls **knockout**, then by the argument above it terminates.  $\square$

Notice that each component of the snapshot object used in our sifter implementation in Figure 1 holds at most 4 identifiers, so it is a  $(4 \cdot \log n)$ -bounded 6-component snapshot object. Combining this with Lemma 17, Lemma 25 and Lemma 26 yields Theorem 6.

## 5 Obstruction-Free Snapshot from Registers

This section establishes Theorem 5. That is, we present an obstruction-free implementation of a  $B$ -bounded  $M$ -component snapshot object from  $M + 1$  registers of size  $\Theta(B + \log n)$ .

Our implementation uses an array  $A[1 \dots M]$  of shared registers and a register  $S$ . Each array entry  $A[i]$  stores a triple  $(w_i, p_i, b_i)$ , where  $w_i \in D$  represents the  $i$ -th entry in the vector  $V$  of the snapshot object,  $p_i$  is a process ID or  $\perp$  which identifies the last process that wrote to  $A[i]$ , and



$b_i \in \{0, 1\}$  is a bounded (modulo 2) sequence number. Initially,  $S = \perp$  and each array entry  $A[i]$  has the value  $(w_i, \perp, 0)$  for some fixed  $w_i \in D$ .

Now suppose process  $p$  calls **update** $(i, x)$ , and this is  $p$ 's  $j$ -th update of the  $i$ -th component of  $V$ . To perform the update,  $p$  first writes its ID to  $S$  and then it writes the triple  $(x, p, j \bmod 2)$  to  $A[i]$ .

To execute a **scan** $()$ , process  $p$  first writes its ID to  $S$ . Then it performs a collect (i.e., it reads all entries of  $A$ ) to obtain a *view*  $a[1 \dots M]$ , and another collect to obtain a second view  $a'[1 \dots M]$ . Finally, the process reads  $S$ . If  $S$  does not contain  $p$ 's ID or if the views  $a$  and  $a'$  obtained in the two collects differ, then  $p$  starts its **scan** $()$  over; otherwise it returns view  $a$ .

Obviously **update** $()$  is wait-free and has step complexity  $O(1)$ . If process  $p$  runs alone for at most  $4m + 3$  steps of its **scan** $()$  operation, it performs a write to  $S$  following by two collects and a read of  $S$ . Since  $p$  runs alone collects are the same and  $p$  reads its own ID from  $S$  and it must terminate. Hence solo step complexity of **scan** $()$  is  $O(M)$ .

To prove linearizability, we use the following linearization points: Each **update** $(i, x)$  operation linearizes at the point when the calling process writes to  $A[i]$ , and each **scan** $()$  operation that terminates linearizes at the point just before the calling process performs its last collect during its **scan** $()$ . (We don't linearize pending **scan** $()$  operations.)

Consider a **scan** $()$  operation by process  $p$  which returns the view  $a = a[1 \dots M]$ . Let  $t$  be the point when that **scan** $()$  linearizes, i.e., just before  $p$  starts its last collect. To prove linearizability it suffices to show that  $A = a$  at point  $t$ .

For the purpose of a contradiction assume that this is not the case, i.e., there is an index  $i \in \{1, \dots, M\}$  such that at time  $t$  the triple stored in  $A[i]$  is not equal to  $a[i]$ . Let  $t_1$  and  $t_2$  be the points in time when  $p$  reads the value  $(w, q, b) = a[i]$  from  $A[i]$  during its penultimate and ultimate collect, respectively. Then  $t_1 < t < t_2$ . Since  $A[i] \neq (w, q, b)$  at time  $t$  but  $A[i] = (w, q, b)$  at times  $t_1$  and  $t_2$ , process  $q$  writes  $(w, q, b)$  to  $A[i]$  at some point in the interval  $(t, t_2) \subseteq (t_1, t_2)$ . Since  $p$  does not write to  $A$  during its **scan** $()$ , this implies  $q \neq p$ .

First suppose  $q$  writes to  $A[i]$  at least twice during  $(t_1, t_2)$ . Each such write must happen during an **update** $()$  operation by  $q$ . Since each **update** $()$  operation starts with a write to  $S$ ,  $q$  writes its ID to  $S$  at least once in  $(t_1, t_2)$ . But since the penultimate collect of  $p$ 's **scan** $()$  starts before  $t_1$  and the ultimate collect finishes after  $t_2$ ,  $S$  cannot change in the interval  $(t_1, t_2)$ , which is a contradiction.

Hence, suppose  $q$  writes to  $A[i]$  exactly once in  $(t_1, t_2)$ ; in particular it writes the triple  $(w, q, b)$  to  $A[i]$  at some point  $t^* \in (t_1, t_2)$ . Recall that each time  $q$  writes to  $A[i]$  it alternates the bit it writes to the third component. Hence, at no point in  $[t_1, t^*]$  the second and third component of  $A[i]$  can have value  $q$  and  $b$ . In particular,  $A[i] \neq (w, q, b)$  at point  $t_1$ , which is a contradiction.

## 6 Obstruction Freedom vs. Randomized Wait-Freedom

In this section, we present a simple technique that transforms any deterministic obstruction-free algorithm into a randomized one that is equally space efficient and is randomized wait-free against the oblivious adversary. Moreover, if the solo step complexity of the deterministic algorithm is  $b$ , then the randomized algorithm guarantees that any process finishes after a number of steps that is bounded by a polynomial function of  $n$  and  $b$ . Precisely, the process finishes in  $O(b(n+b) \log(n/\delta))$  steps, with probability at least  $1 - \delta$ , as stated in Theorem 2.

A naive approach is the following: Whenever a process is about to perform a shared memory step in the algorithm, it can flip a coin, and with probability  $1/2$  it performs the step of the algorithm (called “actual” step), while with the remaining probability it executes a “dummy” step, e.g., reads an arbitrary register. Suppose the solo step complexity of an obstruction-free algorithm is  $b$ . Any

execution of length  $bn$  (i.e., where exactly  $bn$  shared memory steps are performed) must contain a process that executes at least  $b$  steps, and with probability at least  $1/2^{bn}$  that process executes  $b$  actual steps while all other processes execute just dummy steps. Then during an execution of length  $c \cdot b \cdot n \cdot 2^{bn}$  some process runs unobstructed for at least  $b$  actual steps with probability  $1 - 1/e^c$ . Hence, the algorithm is randomized wait-free. This naive transformation yields exponential expected step complexity.

In order to improve the expected step complexity, processes use a biased coin to decide whether to take a larger number of consecutive “dummy” or “actual” steps. Precisely, every process  $p$  tosses a biased coin before its first step, and also again every  $b$  steps. The outcome of each coin toss is heads with probability  $1/n$  and tails with probability  $1 - 1/n$ , independently of other coin tosses. If the outcome of a coin toss by  $p$  is heads, then in its next  $b$  steps,  $p$  executes the next  $b$  steps of the given deterministic algorithm; if the outcome is tails then the next  $b$  steps of  $p$  are *dummy* steps, e.g.,  $p$  repeatedly reads some shared register.

## Proof of Theorem 2

We show that the randomized algorithm described above has the properties specified in Theorem 2.

Let  $\sigma = (\pi_1, \pi_2, \dots)$ , where  $\pi_i \in \mathcal{P}$ , be an arbitrary schedule determining an order in which processes take steps. We assume that  $\sigma$  is fixed before the execution of the algorithm, and in particular before any process tosses a coin. For technical reasons we assume that after a process finishes it does not stop, but it takes *no-op* steps whenever it is its turn to take a step according to  $\sigma$ . Also the process continues to toss a coin every  $b$  (no-op) steps; the outcome of this coin toss has no effect on the execution, but is used in the analysis.

We start with a sketch of the proof. We sort processes by increasing order in which they are scheduled to take their  $(\lambda b)$ -th step in  $\sigma$ , for some  $\lambda = \Theta((n + b) \log(n/\delta))$ . Let  $p_i$  denote the  $i$ -th process in this order. We focus on process  $p_1$  first. We consider  $\lambda$  disjoint *blocks* of  $\sigma$ , where the  $\ell$ -th block, for  $1 \leq \ell \leq \lambda$ , starts with the first step of  $p_1$  after its  $\ell$ -th coin toss, and finishes with the last step of  $p_1$  before its next coin toss. Let  $m_\ell$  denote the number of steps contained in block  $\ell$ ; then  $\sum_\ell m_\ell \leq n\lambda b$  by  $p_1$ 's definition. Further, the number of coin tosses that occur in block  $\ell$  is easily seen to be at most  $O(m_\ell/b + n)$ . These coin tosses, plus at most  $n$  additional coin tosses preceding the block (one by each process), determine which of the steps in the block are actual steps and which are dummy. If all these coin tosses by processes other than  $p_1$  return tails, we say that the block is *unobstructed* (for  $p_1$ ). Such a block does not contain any actual steps by any processes  $p \neq p_1$ . It follows that the probability that block  $\ell$  is unobstructed is at least  $(1 - 1/n)^{O(m_\ell/b + n)}$ . The expected number of unobstructed blocks is then  $\sum_\ell (1 - 1/n)^{O(m_\ell/b + n)}$ , and we show that this is  $\Omega(\lambda)$  using that  $\sum_\ell m_\ell \leq n\lambda b$ . Further, we show that this  $\Omega(\lambda)$  bound on the number of unobstructed blocks holds also with high probability. This would follow easily if for different blocks the events that the blocks are unobstructed were independent; but they are not, as they may depend on the outcome of the same coin toss. Nevertheless the dependence is limited, as each coin toss affects steps in at most  $b$  different blocks and each block is affected by at most  $O(n)$  coin tosses on average. To obtain the desired bound we apply a concentration inequality from [28], which is a refinement of the standard method of bounded differences. Having established that  $\Omega(\lambda)$  blocks are unobstructed, it follows that the probability that  $p_1$ 's coin toss comes up heads at the beginning of at least one unobstructed block is  $1 - (1 - 1/n)^{\Omega(\lambda)} = 1 - e^{-\Omega(\lambda/n)} \geq 1 - \delta/n$  for the right choice of constants. Hence with at least this probability,  $p_1$  finishes after at most  $\lambda b$  steps.

Similar bounds are obtained also for the remaining processes: We use the same approach as above for each  $p_i$ , except that in place of  $\sigma$  we use the schedule  $\sigma_i$  obtained from  $\sigma$  by removing all instances of  $p_j$  except for the first  $\lambda b$  ones, for all  $1 \leq j < i$ . We conclude that with probability

$1 - \delta/n$ ,  $p_i$  finishes after taking at most  $\lambda b$  steps, assuming that each of the processes  $p_1, \dots, p_{i-1}$  also finishes after at most  $\lambda b$  steps. The theorem then follows by applying a union bound.

Next we give the detailed proof. Let  $\lambda = \beta(n + b) \ln(n/\delta)$ , for a constant  $\beta > 0$  to be fixed later. Let  $p_1, \dots, p_k$  be all processes that have at least  $\lambda b$  steps in schedule  $\sigma$ , listed in the order in which they execute their  $(\lambda b)$ -th step. Let  $\sigma_i$ , for  $1 \leq i \leq k$ , be the schedule obtained from  $\sigma$  after removing all instances of  $p_j$  except for the first  $\lambda b$ , for all  $1 \leq j < i$ . For each  $1 \leq i \leq k$ , we identify  $\lambda$  disjoint blocks of  $\sigma_i$ , where for  $1 \leq \ell \leq \lambda$ , the  $\ell$ -th block, denoted  $\sigma_{i,\ell}$ , starts with  $p_i$ 's step following its  $\ell$ -th coin toss, and finishes after the last step of  $p_i$  before its  $(\ell + 1)$ -th coin toss. By  $|\sigma_{i,\ell}|$  we denote the number of steps contained in  $\sigma_{i,\ell}$ . We have  $\sum_{\ell} |\sigma_{i,\ell}| \leq n\lambda b$ , because blocks  $\sigma_{i,1}, \dots, \sigma_{i,\lambda}$  contain in total  $\lambda b$  steps of each of the processes  $p_1, \dots, p_i$ , and fewer than  $\lambda b$  steps of each of the remaining processes.

Observe that if  $p_i$  has not finished before block  $\sigma_{i,\ell}$  begins, and if  $p_i$ 's coin toss before block  $\sigma_{i,\ell}$  returns heads, then  $p_i$  is guaranteed to finish during  $\sigma_{i,\ell}$  if all other steps by non-finished processes during  $\sigma_{i,\ell}$  are dummy steps.

We say that a coin toss *potentially obstructs*  $\sigma_{i,\ell}$  if it is performed by a process  $p \neq p_i$ , and at least one of the  $b$  steps by  $p$  following that coin toss takes place during  $\sigma_{i,\ell}$ . This step will be an actual step only if the coin comes up heads (and  $p$  has not finished yet). We say that block  $\sigma_{i,\ell}$  is *unobstructed* if all coin tosses that potentially obstruct this block yield tails. The number of coin tosses that potentially obstruct  $\sigma_{i,\ell}$  is bounded by  $|\sigma_{i,\ell}|/b + 2n$ , because if process  $p \neq p_i$  takes  $s > 0$  steps in  $\sigma_{i,\ell}$ , then the coin tosses by  $p$  that potentially obstruct  $\sigma_{i,\ell}$  are the at most  $\lceil s/b \rceil$  ones that take place during  $\sigma_{i,\ell}$ , plus at most one before  $\sigma_{i,\ell}$ .

It follows that the probability that  $\sigma_{i,\ell}$  is unobstructed is at least  $(1 - 1/n)^{|\sigma_{i,\ell}|/b + 2n}$ . Thus the expected number of unobstructed blocks among  $\sigma_{i,1}, \dots, \sigma_{i,\lambda}$  is at least  $\sum_{\ell} (1 - 1/n)^{|\sigma_{i,\ell}|/b + 2n}$ . Using now that  $\sum_{\ell} |\sigma_{i,\ell}| \leq n\lambda b$ , and that  $(1 - 1/n)^{x + 2n}$  is a convex function of  $x$ , we obtain that the previous sum is minimized when all  $\lambda$  blocks have the same size, equal to  $nb$ . Thus, the expected number of unobstructed blocks is at least

$$\sum_{1 \leq \ell \leq \lambda} (1 - 1/n)^{|\sigma_{i,\ell}|/b + 2n} \geq \lambda (1 - 1/n)^{(nb)/b + 2n} \geq \lambda (1 - 1/n)^{3n} > \lambda/4^3 = \lambda/64,$$

where for the last inequality we used that  $(1 - 1/n)^n \geq 1/4$ , when  $n \geq 2$ .

Next we use the following result to establish a lower bound on the number of unobstructed blocks with high probability. This result is a special case of [28, Theorem 3.9], which is an extension to the standard method of bounded differences.

**Theorem 27.** *Let  $X_1, \dots, X_{\kappa}$  be independent 0/1 random variables such that  $\Pr(X_j = 1) = \rho$ , for  $1 \leq j \leq \kappa$ . Let  $f$  be a bounded real-valued function defined on  $\{0, 1\}^{\kappa}$ , such that  $|f(x) - f(x')| \leq c_j$ , whenever vectors  $x, x' \in \{0, 1\}^{\kappa}$  differ only in the  $j$ -th coordinate. Then for any  $t > 0$ ,*

$$\Pr(|f(X_1, \dots, X_{\kappa}) - \mathbf{E}[f(X_1, \dots, X_{\kappa})]| \geq t) \leq 2e^{-\frac{t^2}{2\rho \sum_j c_j^2 + 2t \max_j \{c_j\}/3}}.$$

Let the 0/1 random variables  $X_1, X_2, \dots$  denote the outcome of the coin tosses that potentially obstruct at least one of the blocks  $\sigma_{i,1}, \dots, \sigma_{i,\lambda}$ :  $X_j = 1$  if the  $j$ -th of those coin tosses is heads, and  $X_j = 0$  otherwise. Then,  $\Pr(X_j = 1) = 1/n$ . Let  $f(X_1, X_2, \dots)$  be the number of unobstructed blocks. We showed above that  $\mathbf{E}[f(X_1, X_2, \dots)] \geq \lambda/64$ . Further, we observe that flipping the value of  $X_j$  can change the value of  $f$  by at most the number of blocks that  $X_j$  potentially obstructs; let  $c_j$  denote that number. Then,  $\max_j c_j \leq b$ . Finally, since each block  $\sigma_{i,\ell}$  is potentially obstructed by at most  $|\sigma_{i,\ell}|/b + 2n$  coin tosses,

$$\sum_j c_j \leq \sum_{1 \leq \ell \leq \lambda} (|\sigma_{i,\ell}|/b + 2n) = \sum_{1 \leq \ell \leq \lambda} |\sigma_{i,\ell}|/b + 2n\lambda \leq 3n\lambda,$$

Thus,  $\sum_j c_j^2 \leq \sum_j (c_j b) \leq 3n\lambda b$ . Applying now Theorem 27 for  $t = \lambda/128 \leq \mathbf{E}[f(X_1, X_2, \dots)]/2$  gives

$$\Pr(f(X_1, \dots, X_n) \leq t) \leq \Pr(\mathbf{E}[f(X_1, X_2, \dots)] - f(X_1, \dots, X_n) \geq t) \leq 2e^{-\frac{t^2}{6b\lambda + 2tb/3}}.$$

Substituting  $t = \lambda/128$  and  $\lambda = \beta(n + b) \ln(n/\delta)$ , and letting  $\beta = 3(6 \cdot 128^2 + 2 \cdot 128/3)$  yields  $\Pr(f(X_1, \dots, X_n) \leq \lambda/128) \leq 2e^{-3 \ln(n/\delta)} \leq \delta/(2n)$ , for  $n \geq 2$ . Thus, with probability at least  $1 - \delta/(2n)$  at least  $\lambda/128$  of the blocks  $\sigma_{i,1}, \dots, \sigma_{i,\lambda}$  are unobstructed. The probability that  $p_i$  tosses heads before at least one unobstructed block is then at least

$$(1 - \delta/(2n)) \cdot (1 - (1 - 1/n)^{\lambda/128}).$$

Since  $1 - (1 - 1/n)^{\lambda/128} \geq 1 - e^{\lambda/(128n)} > 1 - \delta/(2n)$ , the above probability is at least  $(1 - \delta/(2n))^2 \geq 1 - \delta/n$ .

We have thus far established that for any  $1 \leq i \leq k$ , with probability at least  $1 - \delta/n$  process  $p_i$  finishes after at most  $\lambda b$  steps *under schedule*  $\sigma_i$ . However, schedules  $\sigma$  and  $\sigma_i$  yield identical executions if each of the processes  $p_1, \dots, p_{i-1}$  finishes after executing no more than  $\lambda b$  steps (the executions are identical assuming the same coin tosses in both executions). Then, by the union bound, the probability that all processes  $p_i$  finish after executing no more than  $\lambda b$  steps each is at least  $1 - n \cdot \delta/n = 1 - \delta$ . This concludes the proof of Theorem 2.

## 7 Time and Space Efficient Randomized Test-and-Set

In this section, we present a new randomized TAS algorithm that has the properties stated in Theorem 3. In particular, it uses a logarithmic number of registers, and has almost constant,  $O(\log^* n)$ , expected step complexity against an oblivious adversary. The algorithm combines a known randomized TAS construction [19], with the (deterministic) obstruction-free TAS algorithm from Section 3, which is turned it into a randomized one by applying the technique of Section 6.

We start by observing that since the solo step complexity of the obstruction-free TAS algorithm in Section 3 is  $b = \Theta(\log n)$  (Theorem 1), the technique from Section 6 can be applied. This yields a randomized TAS algorithm that uses  $O(\log n)$  bounded registers, where every process finishes its `compete()` method after at most  $O(n \log^2 n)$  steps, both in expectation and with probability  $1 - O(1/n^c)$ , for any constant  $c > 0$  (by Theorem 2). This step complexity is of course much larger than the nearly constant complexity we want to achieve.

Next we give an overview of the randomized TAS algorithm from [19] that we will use. This algorithm has the desired step complexity, but requires (at least) a linear number of registers rather than logarithmic. To simplify exposition we consider the equivalent weak leader election algorithm rather than the TAS algorithm (see Theorem 7). The algorithm uses a chain of  $n$  sifter objects  $S_1, \dots, S_n$ , alternating with  $n$  splitter objects  $P_1, \dots, P_n$ , and a chain of  $n$  2-process weak leader election objects  $L_n, L_{n-1}, \dots, L_1$ . (A splitter object supports a single operation, `split()`, which returns `win`, `lose`, or `continue`, such that at most one process wins, not all processes lose, and not all continue.)

A process  $p$  starts by invoking the `compete()` method of the first sifter object,  $S_1$ . If  $p$ 's invocation of `compete()` in some sifter  $S_i$  returns `lose`, then  $p$  immediately loses in the weak leader election algorithm. Otherwise, after  $p$  wins in  $S_i$ , it executes the `split()` method of  $P_i$ : if this method returns `lose`, then  $p$  loses immediately, as before; if it returns `continue`,  $p$  invokes the `compete()` method of the next sifter,  $S_{i+1}$ ; while if `split()` returns `win`,  $p$  switches to the chain of 2-process weak leader election objects. In the last case,  $p$  tries to win the 2-process weak leader

elections in  $L_i, L_{i-1}, \dots, L_1$ , in this order. If  $p$  succeeds, it wins the weak leader election algorithm; otherwise it loses, as soon as it loses for the first time in some 2-process weak leader election.

The correctness of the algorithm above follows easily from the next observations. If exactly one process invokes the `split()` method of a splitter  $P_i$ , then this invocation returns `win`, while if there are  $\kappa > 1$  invocations then at least one returns `win` or `continue`, and no more than  $\kappa - 1$  return `continue`. This implies that not all processes lose, and that at most  $n - i + 1$  processes invoke  $P_i$ 's `split()` method, thus no more than  $n$  splitter (or sifter) objects are needed. The `compete()` method of each 2-process weak leader election object  $L_i$  is invoked by no more than two processes: the winner in  $L_{i+1}$  (if it exists), and the at most one winner in  $P_i$ .

In [19], a randomized sifter algorithm is presented that uses  $s \geq 2$  single-bit registers ( $s$  is a parameter), such that the `compete()` method involves just 2 steps, and if at most  $2^s$  processes invoke this method, then at most  $O(s)$  of the invocations return `win`, in expectation. Moreover, for  $s = 2$ , if  $k \geq 2$  invocations of the `compete()` method take place, then the expected number of invocations that return `win` is at most  $k/2 + 1$ . In the following, we will refer to a sifter object implemented by the above algorithm as a *GW-sifter of size  $s$* .

The weak leader election algorithm discussed earlier from [19], uses  $n$  GW-sifters of size  $\log n$  as  $S_1, \dots, S_n$ . This is shown to achieve an expected step complexity of  $O(\log^* n)$ , but requires  $\Theta(n \log n)$  registers in total.

Here we propose instead that different types of sifter objects are used, as follows. The first sifter,  $S_1$ , is a GW-sifter of size  $\log n$ , as before. The next  $\ell = \log^2 \log n$  objects  $S_2, \dots, S_{\ell+1}$ , are GW-sifters of size  $z = 2 \log \log n$ . After that, the next  $m = \beta \log n$  objects  $S_{\ell+2}, \dots, S_{\ell+m+1}$ , for  $\beta > 0$  a sufficiently large constant, are GW-sifters of size 2. Last, sifter  $S_{\ell+m+2}$  is the randomized TAS object obtained by applying Theorem 2 to the deterministic TAS algorithm of Theorem 1, as discussed at the beginning. (Recall that any TAS algorithm is also a 1-sifter.) Objects  $S_i, P_i$ , and  $L_i$ , for  $i > \ell + m + 2$ , are no longer needed.

It is straightforward to verify that this implementation uses  $\Theta(\log n)$  registers: a total of  $\log n + z\ell + 2m = O(\log n)$  registers are used for the sifter objects, and  $O(1)$  registers per object suffice for implementing each splitter  $P_i$  and randomized 2-process weak leader election object  $L_i$  (see [19]).

We compute now the step complexity of the algorithm. For the first sifter, the expected number of invocations that return `win` is  $O(\log n)$ . By Markov's inequality the probability that more than  $2^z = \log^2 n$  invocations return `win` is at most  $O(\log(n)/\log^2 n) = O(1/\log n)$ .

Suppose now that no more than  $2^z$  processes invoke the `compete()` method of the  $z$ -bit sifter  $S_2$  (which happens with probability  $1 - O(1/\log n)$  as argued above). From the analysis in [19] it follows that only the first  $\mu = O(\log^* n)$  of the  $\ell$  sifters  $S_2, \dots, S_{\ell+1}$  are used in expectation. By dividing this sequence of sifters into  $\ell/(2\mu)$  subsequences of  $2\mu$  sifters, and applying Markov's inequality to each, we obtain that the probability all the  $\ell$  sifters are used is  $1/2^{\ell/(2\mu)} = o(1/\log n)$ . Thus, only with probability  $o(1/\log n)$  is sifter  $S_{\ell+2}$  used.

For each sifter  $S_i$ , for  $i \in \{\ell+2, \dots, \ell+m+1\}$ , we have that if  $S_i$ 's `compete()` method is invoked  $\kappa > 1$  times then at most  $\kappa/2 + 1$  invocations return `win` in expectation, thus at most  $\kappa/2$  processes invoke the `compete()` method of the next sifter,  $S_{i+1}$  (because at least one invocation of  $P_i$ 's `split()` does not return `continue`). Then by Markov's inequality, at most  $2\kappa/3$  processes invoke  $S_{i+1}$ 's `compete()` method, with probability at least  $1/4$ . Therefore, no more than  $4 \log_{3/2} n < 8 \log n$  of the  $m$  sifters are used in expectation. By a standard Chernoff bound argument, the probability that the last sifter,  $S_{\ell+m+2}$ , is used can be made smaller than  $n^{-\beta'}$ , for any constant  $\beta' > 0$ , by choosing a sufficiently large constant  $\beta$ .

Combining the above we obtain that the expected number of sifters, other than the last one,

that are used is at most

$$1 + O(\log^* n) + (O(1/\log n) + o(1/\log n)) \cdot m = O(\log^* n),$$

where the first term on the left accounts for  $S_1$ , the second for the expected number of sifters used among the next  $\ell$  sifters, and the third term for the expected number of sifters used among the subsequent  $m$  sifters, where factor  $O(1/\log n) + o(1/\log n)$  is the probability that either more than  $2^z$  processes use  $S_2$  or some process uses  $S_{\ell+2}$ . Using also that the last sifter is used with probability at most  $n^{-\beta'}$  and has step complexity  $O(n \log^2 n)$ , we obtain that the expectation for the maximum number of steps by any process is at most

$$O(\log^* n) + n^{-\beta'} \cdot O(n \log^2 n) = O(\log^* n),$$

for  $\beta' > 1$ . Further the probability that the maximum number of steps is  $O(\log n)$  is  $1 - O(n^{-\beta'})$ , which follows from the probability that the last sifter,  $S_{\ell+m+2}$ , is not used.

## References

- [1] Yehuda Afek, Hagit Attiya, Danny Dolev, Eli Gafni, Michael Merritt, and Nir Shavit. Atomic snapshots of shared memory. *Journal of the ACM*, 40(4):873–890, 1993.
- [2] Yehuda Afek, Eli Gafni, John Tromp, and Paul M. B. Vitányi. Wait-free test-and-set. In *Proceedings of the 6th International Workshop on Distributed Algorithms (WDAG)*, pages 85–94, 1992.
- [3] Zahra Aghazadeh and Philipp Woelfel. Space and time-efficient long-lived test-and-set. In *Proceedings of the 18th International Conference on Principles of Distributed Computing (OPODIS)*, pages 404–419, 2014.
- [4] Dan Alistarh and James Aspnes. Sub-logarithmic test-and-set against a weak adversary. In *Proceedings of the 25th International Symposium on Distributed Computing (DISC)*, pages 97–109, 2011.
- [5] Dan Alistarh, James Aspnes, Keren Censor-Hillel, Seth Gilbert, and Morteza Zadimoghaddam. Optimal-time adaptive strong renaming, with applications to counting. In *Proceedings of the 30th SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, pages 239–248, 2011.
- [6] Dan Alistarh, James Aspnes, Seth Gilbert, and Rachid Guerraoui. The complexity of renaming. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 718–727, 2011.
- [7] Dan Alistarh, Hagit Attiya, Seth Gilbert, Andrei Giurgiu, and Rachid Guerraoui. Fast randomized test-and-set and renaming. In *Proceedings of the 24th International Symposium on Distributed Computing (DISC)*, pages 94–108, 2010.
- [8] James Anderson. Multi-writer composite registers. In *Distributed Computing*, pages 15–30, 1994.
- [9] Hagit Attiya and Arie Fouren. Adaptive and efficient algorithms for lattice agreement and renaming. *Journal on Computing*, 31(2):642–664, February 2002.
- [10] Hagit Attiya, Rachid Guerraoui, Danny Hendler, and Petr Kuznetsov. The complexity of obstruction-free implementations. *Journal of the ACM*, 56(4):24:1–24:33, July 2009.
- [11] Hagit Attiya and Ophir Rachman. Atomic snapshots in  $O(n \log n)$  operations. *Journal on Computing*, 27(2):319–340, April 1998.
- [12] Harry Buhrman, Alessandro Panconesi, Riccardo Silvestri, and Paul Vitányi. On the importance of having an identity or, is consensus really universal? *Distributed Computing*, 18(3):167–176, 2006.
- [13] James Burns and Nancy Lynch. Bounds on shared memory for mutual exclusion. *Information and Computation*, 107(2):171–184, 1993.
- [14] Wayne Eberly, Lisa Higham, and Jolanta Warpechowska-Gruca. Long-lived, fast, wait-free renaming with optimal name space and high throughput. In *Proceedings of the 12th International Symposium on Distributed Computing (DISC)*, pages 149–160, 1998.

- [15] Faith Ellen, Panagiota Fatourou, and Eric Ruppert. Time lower bounds for implementations of multi-writer snapshots. *Journal of the ACM*, 54(6), 2007.
- [16] Faith Ellen Fich, Victor Luchangco, Mark Moir, and Nir Shavit. Obstruction-free algorithms can be practically wait-free. In *Proceedings of the 19th International Symposium on Distributed Computing (DISC)*, pages 78–92, 2005.
- [17] George Giakkoupis, Maryam Helmi, Lisa Higham, and Philipp Woelfel. An  $O(\sqrt{n})$  space bound for obstruction-free leader election. In *Proceedings of the 27th International Symposium on Distributed Computing (DISC)*, pages 46–60, 2013.
- [18] George Giakkoupis, Maryam Helmi, Lisa Higham, and Philipp Woelfel. Test-and-set in optimal space. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 615–623, 2015.
- [19] George Giakkoupis and Philipp Woelfel. On the time and space complexity of randomized test-and-set. In *Proceedings of the 31st SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, pages 19–28, 2012.
- [20] George Giakkoupis and Philipp Woelfel. A tight RMR lower bound for randomized mutual exclusion. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 983–1002, 2012.
- [21] Wojciech Golab, Danny Hendler, and Philipp Woelfel. An  $O(1)$  RMRs leader election algorithm. *SIAM Journal on Computing*, 39(7):2726–2760, 2010.
- [22] Rachid Guerraoui, Maurice Herlihy, and Bastian Pochon. Toward a theory of transactional contention managers. In *Proceedings of the 24th SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, pages 258–264, 2005.
- [23] Maurice Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124–149, 1991.
- [24] Maurice Herlihy, Victor Luchangco, and Mark Moir. Obstruction-free synchronization: Double-ended queues as an example. In *Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS)*, pages 522–529, 2003.
- [25] Maurice Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, 1990.
- [26] Prasad Jayanti, King Tan, and Sam Toueg. Time and space lower bounds for nonblocking implementations. *SIAM Journal on Computing*, 30(2):438–456, 2000.
- [27] Clyde Kruskal, Larry Rudolph, and Marc Snir. Efficient synchronization on multiprocessors with shared memory. *ACM Transactions on Programming Languages and Systems*, 10(4):579–601, 1988.
- [28] Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer-Verlag, 1998.



- [29] Alessandro Panconesi, Marina Papatriantaflou, Philippas Tsigas, and Paul M. B. Vitányi. Randomized naming using wait-free shared variables. *Distributed Computing*, 11(3):113–124, 1998.
- [30] Eugene Styer and Gary Peterson. Tight bounds for shared memory symmetric mutual exclusion problems. In *Proceedings of the 8th SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, pages 177–191, 1989.
- [31] John Tromp and Paul Vitányi. Randomized two-process wait-free test-and-set. *Distributed Computing*, 15(3):127–135, 2002.